Virus Evolution, 2017, 3(2): vex025

doi: 10.1093/ve/vex025
Resources

Scalable relaxed clock phylogenetic dating

E. M. Volz[1,*,†] and S. D. W. Frost[2]

[1]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, Norfolk Place, London W2 1PG, UK and [2]Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge, CB3 0ES, UK

*Corresponding author: E-mail: e.volz@imperial.ac.uk

†http://orcid.org/0000-0001-6268-8937

Abstract

Molecular clock models relate observed genetic diversity to calendar time, enabling estimation of times of common ancestry. Many large datasets of fast-evolving viruses are not well fitted by molecular clock models that assume a constant substitution rate through time, and more flexible relaxed clock models are required for robust inference of rates and dates. Estimation of relaxed molecular clocks using Bayesian Markov chain Monte Carlo is computationally expensive and may not scale well to large datasets. We build on recent advances in maximum likelihood and least-squares phylogenetic and molecular clock dating methods to develop a fast relaxed-clock method based on a Gamma-Poisson mixture model of substitution rates. This method estimates a distinct substitution rate for every lineage in the phylogeny while being scalable to large phylogenies. Unknown lineage sample dates can be estimated as well as unknown root position. We estimate confidence intervals for rates, dates, and tip dates using parametric and non-parametric bootstrap approaches. This method is implemented as an open-source R package, *treedater*.

Key words: molecular clock; relaxed clock; Ebola.

1. Introduction

Pathogen sequence data can provide important information about the timing and spread of infectious diseases, particularly for rapidly evolving pathogens such as RNA viruses. Such pathogens have been dubbed 'measurably evolving' (Drummond et al. 2003), as sequences typically accumulate mutations over epidemiological timescales of years or even months. By using sampling dates in conjunction with sequence data, it is possible to estimate the rate of evolution, and hence generate phylogenetic trees calibrated in calendar time. These 'time-trees' are more straightforward to interpret in terms of the time to the most recent common ancestor and changes in effective population size, which can then be linked to external epidemiological information, as in the case of the spread of hepatitis C virus in Egypt during antischistosomiasis injection campaigns (Pybus et al. 2003) and the spatial spread of rabies virus in raccoons in the USA (Biek et al. 2007).

While there may be a fairly constant average rate of evolution over epidemiological timescales, there may be variation in evolutionary rates across lineages in the phylogenetic tree; failure to account for this variation may lead to incorrect inferences of evolutionary rates and dates. This has led to the development of computationally-intensive Bayesian approaches, which assume an underlying model for how evolutionary rates vary across the phylogeny (Drummond et al. 2006).

With the growth in the size of pathogen sequence datasets, it is becoming increasingly difficult to apply Bayesian relaxed-clock methods. There have been several recent developments in fast, approximate methods for generating time-trees from sequence data (To et al. 2015; Jones and Poon 2016); however, these approaches do not flexibly model rate variation, which can affect estimates of the evolutionary rate (Duchêne et al. 2016).

We present an approach to fit a relaxed clock to a non-clocklike phylogenetic tree with associated data on sampling times. Using simulated data, we demonstrate that explicit incorporation of a relaxed clock leads to more accurate inference of the mean rate of evolution in addition to providing information on the variation in evolutionary rates. Our implementation

generates confidence intervals for the evolutionary rate and the time to the most recent common ancestor using parametric bootstrapping (PB), which lends itself well to parallelization. Our approach allows testing of a relaxed (vs. a strict) molecular clock, as advised by Duchene et al. (2015); it can detect outlier lineages associated with unusually high or low rates of evolution, and it can infer missing sampling times. We demonstrate these features using a large ($n = 1{,}610$) genome-scale dataset of Ebola virus sequences from the West African Ebola epidemic (Dudas et al. 2015) and compare the performance of the new method with other state-of-the-art methods.

2. Methods

Given a lineage i in a binary phylogeny $\mathcal{T}$, we define ω_i to be the rate of evolution in units of substitutions per site per unit time along lineage i. The data takes the form of a branch lengths b_i in $\mathcal{T}$ with units of substitutions per site, which can be estimated from a sequence alignment using maximum likelihood (ML), a Bayesian approach, or a distance-based approach such as neighbour joining.

We assume that the length of the sequence alignment, denoted S, is known. If S is large and ω_i is sufficiently small, the probability of reversions on lineage i will be small and the number of substitutions on lineage i is well approximated by a Poisson distribution; this approximation is also known as the Langley–Fitch model (Langley and Fitch 1974). We denote the actual number of substitutions on branch i as $s_i = Sb_i$, the temporal length of lineage i as $\tau_i = t_i - t_{a(i)}$, where t_i is the time of the $i'th$ node descended from lineage i, $a(i)$ is the node from which lineage i is descended, and $t_{a(i)}$ is the time of the ancestor of node i. We model substitutions as arising from a Poisson process with branch-specific rate $\lambda_i = \tau_i\omega_i$, such that $s_i \sim \text{Pois}(\lambda_i)$.

To account for rate variation, λ_i is modeled using a Gamma distribution, where the variance of the rates depends upon the branch lengths: $\lambda_i \sim \Gamma(r, \varphi\tau_i)$, where r and φ are shape and scale parameters to be estimated. With the Gamma-Poisson mixture so defined, the distribution of $s_i|r, \varphi$ is negative binomial (Greenwood and Yule 1920):

$$s_i|r, \varphi \sim NB\left(r, \frac{\varphi\tau_i}{1+\varphi\tau_i}\right). \tag{1}$$

Given proposed values of r, φ, and τ_i, it is possible to compute the most probable value of λ_i and by extension the branch rate ω_i.

$$\begin{aligned} l(\lambda_i|\tau_i, r, \varphi) &= \log(p(s_i|\lambda_i)p(\lambda_i|r, \varphi)) \\ &= c + s_i\log(\lambda_i) - \lambda_i + (r-1)\log(s_i) - \frac{s_i}{\varphi\tau_i}, \end{aligned} \tag{2}$$

where c is a constant independent of λ. This likelihood is a convex function of λ_i and has a unique optimum at

$$\lambda_i^* = \frac{\varphi\tau_i}{\varphi\tau_i + 1}(s_i + r - 1). \tag{3}$$

The conditional ML estimate of ω_i is then given by $\omega_i^* = \lambda_i^*/(\tau_i S)$.

It remains to develop a strategy for jointly optimising the likelihood of all node dates t_i, Gamma parameters r and φ, and the position of the root of the phylogeny. For now, let us assume that the data take the form of a bifurcating rooted phylogeny with branch lengths in units of substitutions per site and that all tip dates are known. We will subsequently relax the assumption that the input tree is rooted and that all tip dates are known. Following the convention in (To et al. 2015), we index nodes such that $i = 1, 2, \cdots, n-1$ correspond to internal nodes and indices $i = n, \cdots, 2n-1$ correspond to tip dates. Given a proposal of internal node times $(t_i)_{i=1:n-1}$ a conditional ML estimate of (r, φ) is found by optimising the log likelihood

$$l(r, \varphi|(t_i)_{i=1:n-1}) = \sum_{i=1:n-2} log\left(f_{nb}\left(\left(s_i|r, \frac{\varphi\tau_i}{1+\varphi\tau_i}\right)\right)\right), \tag{4}$$

where $f_{nb}(\cdot)$ is the negative binomial density. Optimising this likelihood is straightforward using gradient-descent or simplex strategies; the computational cost of a likelihood computation is linear in sample size.

It is also straightforward to obtain a very good approximation to the ML $(t_i)_{i=1:n-1}$ conditional on (r, φ) and $(\omega_i)_{i=1:n-1}$ using the least-squares approach described in (To et al. 2015). Briefly, we minimize the weighted residual sum of squares

$$RSS((t_i)_{i=1:n-1}|(\omega_i)_{i=1:n-1}) = \sum_{i=2}^{2n-1} \frac{1}{\sigma_i^2}(b_i - \omega_i\tau_i)^2 \tag{5}$$

The parameter σ_i is an approximation to the variance of b_i. Following the approach in (To et al. 2015), we use $\sigma_i = (b_i s + c)/s$, where c is a tuneable parameter and s is the sequence length.

Minimizing RSS is linear in sample size. It is also possible to solve a constrained least-squares problem if we wish to enforce the constraint that $t_i > t_{a(i)}$. This is implemented using the quadratic-programming algorithms implemented in the *mgcv* and *quadprog* R packages (Wood 2006; Turlach and Weingessel 2013). The main difference between this optimization and the one described in (To et al. 2015) is that branch-specific ω_i take the place of a constant substitution rate ω.

Whereas individually optimising ω_i or (r, φ) or $(t_i)_{i=1:n-1}$ is straightforward conditional on other parameters, rapid optimization of all parameters r, φ, and $(t_i)_{i=1:n-1}$ is challenging. We therefore adopt a fast heuristic iterative approach described in algorithm 1.

Algorithm 1: The *treedater* algorithm given a rooted tree and tip dates.

Data: A rooted binary tree $\mathcal{T}$ with branches in units of substitutions per site and dates for all sampled lineages; a tolerance threshold $d\mathcal{L}'$ for convergence ;
Result: A rooted tree $\mathcal{P}$ with branches in units of time; estimate of mean substitution rate; estimate of substitution rate on each branch; parameters of relaxed clock model;
Initialize index $k \leftarrow 0$;
Initialize mean substitution rate $\omega^{(k)}$ by root-to-tip regression [5, 26] ;
Initialize $(\omega_i)_{i=1:n-1}^{(k)}$ to $\omega^{(k)}$;
Initialize $r^{(k)} = 3$ and $\phi^{(k)}$ to correspond to $\omega^{(1)}$ with moderate rate variation (coefficient of variation of rates$\approx .5$) ;
Initialize log likelihood $\mathcal{L}^{(k)} = -\infty$ and $d\mathcal{L} = \infty$;
repeat
  Compute $(t_i)_{i=1:n-1}^{(k+1)}|(\omega_i)_{i=1:n-1}^{(k)}$ using equation 5 ;
  Compute $r^{(k+1)}$ and $\phi^{(k+1)}|(t_i)_{i=1:n-1}^{(k+1)}$ by maximizing equation 4 ;
  Compute $(\omega_i)_{i=1:n-1}^{(k+1)}$ using equation 3 ;
  Compute $\mathcal{L}^{(k+1)}$ using equation 4 and $d\mathcal{L} \leftarrow \mathcal{L}^{(k+1)} - \mathcal{L}^{(k)}$;
  $k \leftarrow k + 1$;
until $d\mathcal{L} < d\mathcal{L}'$;

This algorithm can be repeated for multiple starting conditions of the initial substitution rate to improve the quality of the estimate.

2.1 Estimating root position

Given a rooted tree $\mathcal{T}$ with branches in units of substitutions, x_i will denote the root-to-tip (RTT) distance for sampled lineage i; this is the sum of all branch lengths between the root node and tip i. A common approach to rate estimation is to regress x_i on the known date of sampling t_i. The slope of the regression line is an estimate of the mean rate of substitution per unit time where the correlation due to shared ancestry has been neglected. This approach is implemented in the software *TempEst* (Rambaut et al. 2016) and in the *ape* R package (Paradis, Claude, and Strimmer, 2004).

RTT regression also provides a fast means of optimising root position given an unrooted tree $\mathcal{T}'$. We adapt the approach implemented by the *rtt* function which is part of the *ape* R package (McCloskey 2015). In brief, given a proposed root edge *u*, the residual sum of squares of the RTT regression using the tree rooted on *u* can be computed. This can be repeated for every branch in the tree. The *treedater* algorithm uses this heuristic approach to identify a set of n_r good candidates for the root position. Algorithm 1 can be repeated for every good candidate root position and the dated tree with the highest likelihood is returned. The complete algorithm is described in algorithm box 2.

Algorithm 2: The *treedater* algorithm given a rooted tree and tip dates.

Data: An unrooted tree $\mathcal{T}'$ with branches in units of substitutions per site and dates for all sampled lineages; a tolerance threshold $d\mathcal{L}'$ for convergence ; parameter n_r = number of root positions to attempt ;
Result: A rooted tree $\mathcal{P}$ with branches in units of time; estimate of the mean substitution rate; estimate of substitution rates on each branch; parameters of relaxed clock model;
for *each lineage* u **do**
  Create a root node a on u at position x_a chosen to minimize RSS of root-to-tip regression ;
  Record RSS_u ;
end
Sort edges u in order of increasing RSS_u giving vector $(u)_{i=1:n-2}$;
for $i = 1$ **to** n_r **do**
  Apply algorithm 1 to tree rooted on u_i yielding estimated time tree $\mathcal{P}_i$;
  Record likelihood $\mathcal{L}_i$ and $\mathcal{P}_i$;
end
Compute $i^* = \text{argmax}_i(\mathcal{L}_i)$;
return $\mathcal{P}_{i^*}$;

2.2 Estimating tip dates

In many real applications, dates of lineage sampling may not be known with certainty. Sometimes, the exact sampling time is not known; it may be missing from the annotations, or recorded to a particular precision (e.g. the calendar year rather than the date). Given an initial guess of tip dates $(t_i^{(0)})_{i=n:2n-1}$ and lower bounds $(l_i^{(0)})_{i=n:2n-1}$ and upper bounds $(u_i^{(0)})_{i=n:2n-1}$ we can modify algorithm 1 to optimize tip dates in each iteration. At step *k* of algorithm 1, we model the number of substitutions on tip i as Poisson with rate $\lambda_i(t_i) = (t_i - t_{a(i)})\omega_i S$. We then optimize the log likelihood

$$l(t_i^{(k+1)} | t_{a(i)}^{(k+1)}, \omega_i) = \log(f_{\text{Poisson}}(s_i | \lambda_i(t_i))). \quad (6)$$

This univariate optimization is then repeated for each uncertain tip date t_i at each iteration *k*. Note that this is a heuristic optimization and in general will not return the unique optimal combination of tip and internal node dates. Better optima can be found by repeating the *treedater* algorithm with different guesses of the initial tip dates. The performance of this tip-dating variant of the *treedater* algorithm is explored in simulation results below.

2.3 Parametric bootstrap

Because the likelihood under the *treedater* model is optimized heuristically, it is challenging to apply standard likelihood based approaches such as profiling to estimate confidence intervals. A standard approach for assessing uncertainty in phylogenetic analyses is to perform non-PB, in which columns in the multiple sequence alignment are resampled with replacement in order to generate new datasets. Fast approximations to non-PB for phylogenetic reconstruction have also been proposed (Nguyen et al. 2015), and the latest version of the least squares dating (LSD) software also includes PB routines (To et al. 2015) (accessed 6 April 2017). In addition to running *treedater* on multiple bootstrapped phylogenies, Monte Carlo simulation and PB approaches offer a highly flexible and parallelizable approach for estimating uncertainty in substitution rates and node dates (Efron and Tibshirani 1994). The PB approach implemented in *treedater* assumes 1, the data were generated under the strict or relaxed clock model as implemented in *treedater*, so that substitutions on each branch will follow a NB distribution as in equation 1. 2, The sampling distribution of estimated rates and time of the most recent common ancestors (TMRCAs) is asymptotically normal and the SD of the sampling distribution is well approximated by the PB distribution of estimated rates and TMRCAs.

The *treedater* PB algorithm works by simulating n_{pb} synthetic datasets $\tilde{\mathcal{T}}_{j=1:n_{pb}}$. These are generated by simulating a tree (rooted or unrooted) with identical topology as the original data but with branch lengths distributed:

$$\tilde{b}_i \sim \text{NB}\left(\hat{r}, \frac{\hat{\phi}\hat{\tau}_i}{1+\hat{\phi}\hat{\tau}_i}\right)/S \quad (7)$$

where parameters $\hat{\cdot}$ correspond to the pseudo-ML estimate provided by algorithms 1 or 2.

The *treedater* algorithm (1 or 2) is applied to each $\tilde{\mathcal{T}}_j$ providing a Monte Carlo sample of substitution rates $\tilde{\omega}_{j=1:n_{pb}}$ and other parameter $\tilde{\varphi}_{j=1:n_{pb}}$ and $\tilde{r}_{j=1:n_{pb}}$ and node dates $\tilde{t}_{ij}$. The estimate of the sampling SD of model parameters is the SD of the PB sample.

2.4 Detecting outliers

The *treedater* algorithm provides several statistics associated with each sampled lineage that can be useful for identifying outlier lineages; these may represent sequencing error or samples that are poorly described by the fitted substitution model. In such cases, outliers can be identified and removed in order to produce a data set that the given molecular clock model can better fit. Existing software, such as *TempEst* (Rambaut et al. 2016), uses RTT regression in order to perform these comparisons.

For each sampled lineage *treedater* provides 1, the estimated log likelihood of the branch length under the substitution model; 2, the estimated substitution rate for that branch; 3, a *P*-value for the branch length under the fitted substitution model; and 4, a q-value

(Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001), which provides a quantitative measure of the extent to which the lineage is an outlier under the fitted model and adjusts for multiple testing bias. *treedater* uses q-values computed using the *p.adjust* method in R (R Core Team 2016). Lineages may be identified and excluded as outliers if their q-value is less than a user-defined threshold q^*; the proportion of outliers detected that are expected to be false-discoveries (not true outliers) is q^*.

2.5 Statistical test for detecting a relaxed molecular clock

A strict clock is unlikely to hold in principle; in practice, however, there may be insufficient information in order to fit a relaxed clock. Fitting a relaxed clock in this case may risk overfitting the data (Duchêne et al. 2016). We propose a simple frequentist test to reject the null hypothesis of a strict clock by computing the null distribution of the coefficient of variation (CV) of rates across the tree.

The test utilizes the PB described in Section 2.3 to produce a distribution of CV under the null (Stute, Manteiga, and Quindimil 1993). First, the *treedater* algorithm is fit to the data under a strict clock (Poisson substitution model). Then, the PB from Section 2.3 is applied using a relaxed molecular clock. This provides a bootstrap distribution of estimated CV of rates under the null hypothesis that the clock is strict. Finally, the relaxed clock model is fitted to the original data set and the CV is estimated. If the CV under the relaxed clock falls outside a prespecified quantile of the bootstrap distribution, the null is rejected.

2.6 Simulations

To compare the performance of *treedater* with other dating methods, we use simulations from two recent publications (To et al. 2015; Jones and Poon 2016). The reader is referred to the original publications for a detailed description of simulation design; brief descriptions of the simulated datasets are as follows. In To et al. (2015), simulations are developed using both strict and relaxed clock models corresponding to HIV transmission chains, which presents a challenging scenario for molecular clock dating. Simulated data including BEAST configuration files are available at https://github.com/emvolz-phylodynamics/treedater-simulation-experiments. Four scenarios are developed corresponding to different distributions of sample dates through time and different levels of within-host genetic diversity. We use unrooted phylogenies estimated by ML using PhyML, which were previously computed by To et al. (2015).

Jones and Poon (2016)conducted a birth-death simulation to generate a genealogy and assume a strict molecular clock.

To evaluate the performance of the different methods, we use several statistics. To measure precision, we define the relative root mean square error of substitution rates to be (RRMSEω) to be

$$\mathrm{RRMSE}_\omega = \frac{1}{\omega}\sqrt{\langle \omega - \hat{\omega}_k \rangle}$$

where index k denotes the simulation replicate and brackets denote arithmetic mean. To measure bias of estimated rates we define the relative mean error

$$\mathrm{RME}_\omega = \frac{1}{\omega}\langle \omega - \hat{\omega}_k \rangle.$$

Similarly, for estimated TMRCAs we define

$$\mathrm{RMSE}_t = \sqrt{\langle t_{mrca} - \hat{t}_{mrca_k} \rangle}$$

and

$$\mathrm{ME}_t = \langle t_{mrca} - \hat{t}_{mrca_k} \rangle.$$

For simulations in (To et al. 2015), *treedater* is compared to the following other methods:

- The QPD least-squares dating algorithm (To et al. 2015) with temporal constraints ($t_i > t_{a(i)}$).
- Bayesian relaxed molecular clock with estimated topology using BEAST (Drummond et al. 2006).
- RTT regression (Drummond et al. 2003; Rambaut et al. 2016).

All methods except for BEAST use an unrooted ML input tree estimated using PhyML (Guindon et al. 2010). Note that BEAST is a complex Bayesian method with many tuneable parameters. Bayesian prior distributions used to generate BEAST estimates closely mirror how the data was simulated (To et al. 2015); however, in To et al. (2015), performance of BEAST was not optimized with respect to all available parameters. A Uniform(0,1) prior was used for the molecular clock rate which is not standard in BEAST. Furthermore, the coalescent tree prior was fixed at a constant size. To improve on performance of BEAST reported in To et al. (2015), we re-ran BEAST using a flexible skyride coalescent prior (Minin, Bloomquist, and Suchard 2008) and with longer MCMC chain length (50 million iterations). We ensured that effective sample size of all parameters exceeded one thousand. For comparisons with RTT and QPD, we re-use data from a previous publication (To et al. 2015).

2.7 West African Ebola epidemic

As an additional test of our approach, we fitted our model to sequence data from the West African Ebola epidemic (2013–2016). Near-full length genomes ($n = 1{,}610$) of Zaire Ebola virus from Africa, sampled between 17 March 2014 and 24 October 2015 have been collated, processed, and analysed using BEAST by Dudas et al. (2017) and have been shared by the authors under a Creative Commons 4.0 license at http://github.com/ebov/space-time. The sequence alignment was extracted from the BEAST XML file using BEASTgen v1.0.2, and we estimated a ML tree using IQTREE v.1.5.3 (Nguyen et al. 2015) using an HKY + F+G4 model applied to each of four partitions (first, second, and third codon positions, plus the non-coding region), the same underlying model used in the BEAST analysis of Dudas et al., chosen so as to maximize the comparability between the different approaches. An initial tree was generated using default options, then refined using a more thorough nearest-neighbor interchange search. Sample collection dates (or imputed dates) were also provided by Dudas et al. We ran *treedater* using the top 10 root positions identified using RTT regression, with two starting values for the evolutionary rate. Results from *treedater* were compared to those from the QPD least-squares dating algorithm (To et al. 2015), and from the maximum clade credibility tree and a sample of 1,000 trees from the posterior distribution from the analysis of Dudas et al., obtained using a relaxed molecular clock. Inference of the cumulative number of infected individuals from time-calibrated trees was performed using skyspline (Volz, Romero-Severson, and Leitner 2017), assuming a 15-day infectious period, and compared to the cumulative number of

reported cases from Guinea, Sierra Leone and Liberia, as collated by the WHO and processed by the CDC (https://www.cdc.gov/vhf/ebola/csv/graph1-cumulative-reported-cases-all.xlsx).

3. Results

The *treedater* algorithm provides robust estimates of substitution rates and node dates across a range of simulation scenarios presented by To et al. (2015) and Jones and Poon (2016), which includes a range of sample designs and strict or relaxed molecular clocks. Figure 1 illustrates estimates from *treedater* in a relaxed clock simulation scenario from (To et al. 2015) where *treedater* performed the best in comparison to three other methods (BRMC, Drummond et al. 2006; lsd-QPD, To et al. 2015; and RTT, Rambaut et al. (2016). In this scenario (D750_11_10), *treedater* provides accurate and precise estimates of the mean substitution rate, as well as good coverage of estimated rates and lineages through time. In comparison to other simulation scenarios, this scenario was characterized by relatively large sample size ($n = 110$), a balanced tree topology, and samples distributed throughout the history of the tree (some samples near root of tree). Note that all methods performed well or poorly in at least one scenario and estimated substitution rates for all scenarios are illustrated in supporting Supplementary Figure S2.

The bias and error of estimated substitution rates and TMRCA using *treedater* in comparison to these other methods is tabulated for four relaxed clock simulation scenarios in Table 1 and in supporting Supplementary Figures S1, S2, and S2. Among the four performance metrics and four scenarios, *treedater* provides the best performance in 9 out of 16 comparisons with BEAST, QPD, and RTT. For metrics and scenarios where *treedater* was not the best performing method, it was usually the second best performing method by a small margin (results not shown).

In most scenarios, confidence intervals estimated using the PB provided good coverage of the true values, however in scenario D995_11_10, coverage of the estimated TMRCA fell to 66 per cent. In comparison to other simulation scenarios, this scenario was characterized by an imbalanced ladder-like topology with many samples near the root of the tree. Nevertheless, the mean error of the estimated TRMCA in this scenario was quite small and the best performing of the four methods compared.

For 50 strict clock birth-death simulations in Jones and Poon (2016), we find a weighted RMSE of 21.22 for the estimated TMRCA. This can be compared to values in the study by Jones and Poon (2016) of 22.1 for the *node.dating* method with 10^4 steps and 20.1 for BEAST with correctly specified priors and a strict clock (SMC). Note that in Jones and Poon (2016), the *node.dating* algorithm is run with a fixed root position determined by RTT, whereas with *treedater,* the root position was optimized among 10 candidate branches, which may partially explain the difference in performance. Among these methods (*treedater, node.dating*, BEAST SMC), *treedater* is by far the fastest method with a mean runtime of 1.16 seconds, which can be compared to 5,950 seconds for *node.dating* with 10^4 steps or 6,840 seconds for BEAST SMC with 10^6 steps (Jones and Poon 2016).

3.1 Testing a relaxed clock versus a strict clock

We applied our relaxed clock test to the simulated data from To et al. (2015) including trees generated under strict and relaxed clocks. Across 400 simulations and four scenarios we find that the test has a 100 per cent true positive rate for detecting the relaxed clock. On the other hand, across 400 strict clock simulations, we find a false-positive rate of 34.8 per cent (erroneous detection of a relaxed clock). The majority of the false positives (89 of 139) were concentrated on a single scenario (D750_3_25). This scenario was characterized by relatively small sample size ($n = 75$) and a distant TMRCA well before the earliest sample.

We also applied the relaxed clock test to the strict clock birth-death simulations in Jones and Poon (2016). In 49 of 50 simulations, the relaxed clock test correctly failed to reject the strict clock null hypothesis (false-positive rate = 2%).

3.2 Inference with uncertain times of sampling

We evaluated *treedater* in the presence of uncertain times of sampling ('tip dates') by modifying simulated trees from To et al. (2015). We randomly selected 20 per cent of sampled lineages and treated their tip date as missing data. The starting conditions for uncertain tip dates were drawn from a uniform distribution spanning the range of all non-missing tip dates. Note that this simulation scenario represents extreme uncertainty in tip dates; in most real-world situations, some prior information would be available that would allow stronger constraints to be placed on unknown tip dates (e.g. nearest week or month of sampling for pathogen sequence data). Figure 2 shows the residuals of estimated tip dates. In this scenario, the *treedater* algorithm estimates tip dates along with other node dates and parameters with little bias in tip dates (MRE = 12.4%). Relative to the starting conditions, RMSE of estimated tip dates was 59 per cent lower. Estimation of molecular clock parameters is deteriorated by missing tip dates in this extreme scenario; the RRMSE of the mean substitution rate across all scenarios is 32.5% (compare to Table 1).

3.3 West African Ebola virus epidemic

In addition to simulated data, we also analysed a large sequence dataset from the West African Ebola virus epidemic, collated, processed and analysed previously by Dudas et al. (2017). The dataset is composed of many ($n = 1{,}610$) near-full-length genome sequences, many—but not all—of which have sampling dates, as opposed to sampling months or years; in the analysis of Dudas et al., 29 collection dates were imputed. The dataset was also cleaned by removing potential T-to-C hypermutations that may have arisen through ADAR editing, and by the removal of sequences sampled from re-emerged transmission chains originating from individuals with persistent Ebola virus infection (Blackley et al. 2016). The latter are associated with low genetic divergence, consistent with a reduced evolutionary rate in persistently infected individuals (Holmes et al. 2016). As such, this dataset has been curated but still presents a challenge for phylogenetic dating due to the large sample size and relatively short sampling time frame. There are also external epidemiological data on the timing and the dynamics of the epidemic that can be used to validate inferences from sequence data alone.

The first documented cases of Ebola virus infection in humans occurred in Guinea in December 2013, hence the time of the most recent common ancestor of the sequences is likely to be no earlier than this (Table 2). Using a sample of 1,000 phylogenies from the BEAST fits obtained by Dudas et al., we computed the posterior distribution of the time of the most recent common ancestor. The mean and median TMRCAs were 7 December 2013 and 13 December 2013, respectively, with a 95 per cent credible interval of 13 September 2013 to 26 January

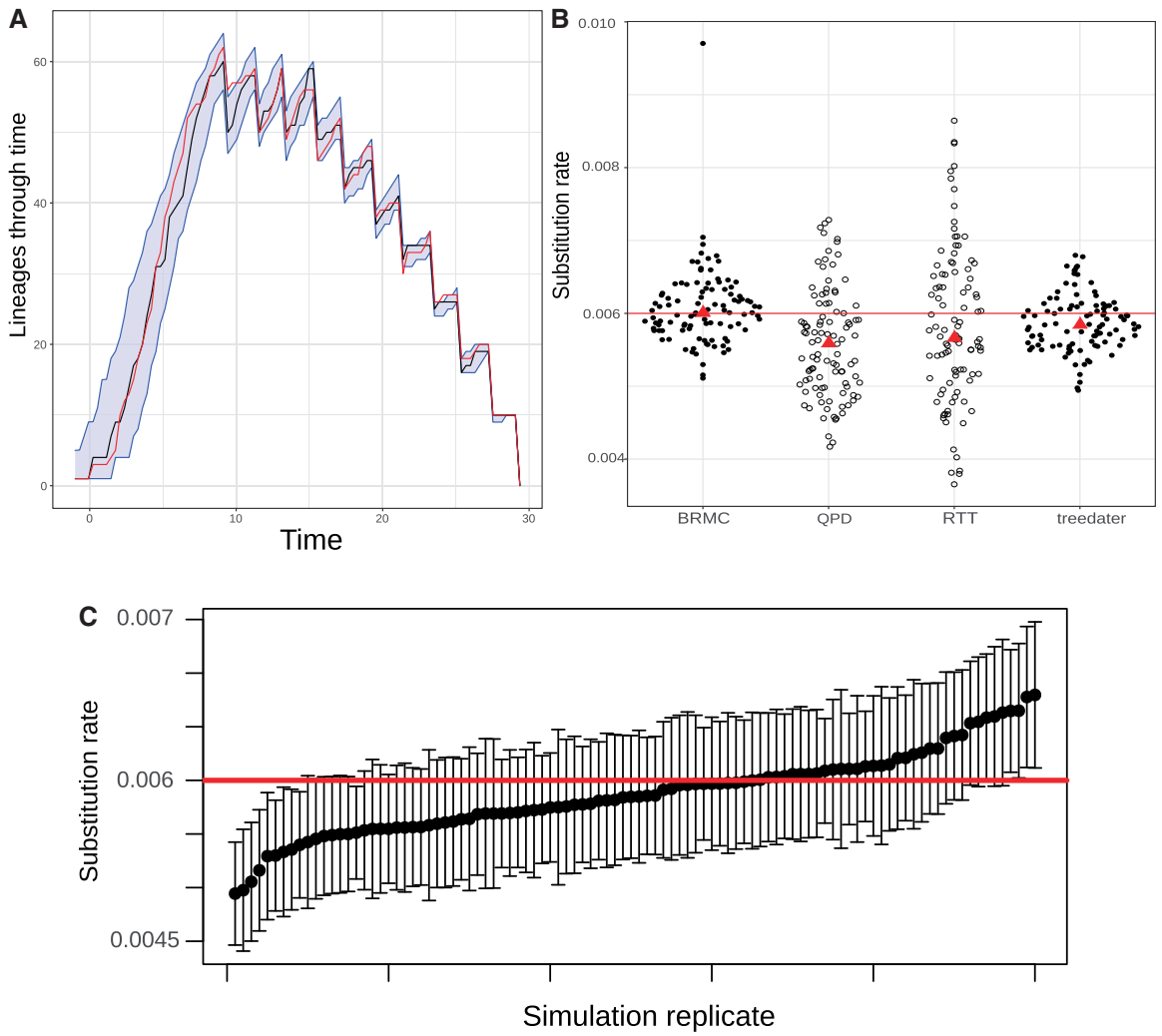

Figure 1. Evaluating performance of *treedater* using simulations from To et al. (2015) under a relaxed clock model. This simulation corresponds to scenario D750_11_10 in which *treedater* has similar performance as BEAST. A, Estimated (black) and actual (red) lineages through time for a single randomly chosen simulation replicate. The shaded region shows estimated confidence intervals. Note that lineages increase during sampling events, producing the jagged pattern. B, Estimates of the mean substitution rate obtained by three methods compared in To et al. (2015) in addition to *treedater* over 100 simulation replicates. The red line indicates the true value. The red triangle indicates the mean estimate for each method across all simulation replicates. Note that estimates with RTT and QPD were recycled from an earlier publication (To et al. 2015) and are shown with unfilled points. C, Estimated mean substitution rate for each simulation replicate using *treedater* and estimated confidence intervals. The red line indicates the true value.

Table 1. Bias and precision of *treedater* algorithm over four scenarios and 100 relaxed clock simulations from To et al. (2015).

Scenario	RMEω	RRMSEω	ME$_{tmrca}$	RMSE$_{tmrca}$	Coverageω	Coverage$_{tmrca}$
D750_11_10	−0.021 (−0.005)	0.068 (0.064)	**−0.023 (−0.082)**	**0.097 (0.164)**	0.86	0.85
D750_3_25	**−0.022 (0.032)**	0.133 (0.129)	**−0.043 (−0.1344)**	**0.184 (0.194)**	0.84	0.90
D995_11_10	0.031 (−0.023)	**0.097 (0.115)**	**0.012 (−0.012)**	0.032 (0.028)	0.88	0.66
D995_3_25	**−0.012 (0.026)**	**0.121 (0.140)**	−0.009 (−0.002)	0.067 (0.058)	0.87	0.84

ω is the mean substitution rate and 'Coverage' refers to frequency with which 95 per cent confidence intervals covered the true value. In parentheses are shown the best performance measures in a pooled comparison of RTT, least squares dating, least squares dating QPD, and BEAST relaxed molecular clock models. Metrics for which *treedater* was the best performing algorithm are shown in bold face.

2014, with the TMRCA of the maximum clade credibility tree of 5 December 2013. Using a ML tree, both RTT regression and *node.dating* gave estimate of the TMRCA (1 November 2013 and 31 October 2013) that were much earlier than the first documented case in humans. In contrast, the point estimate of the TMRCA using *treedater* was within a few days of that inferred from the BEAST maximum clade credibility tree. We detected substantial rate variation in this data set, which may explain the discordant results between methods that explicitly account for rate variation (BEAST and *treedater*) and other methods.

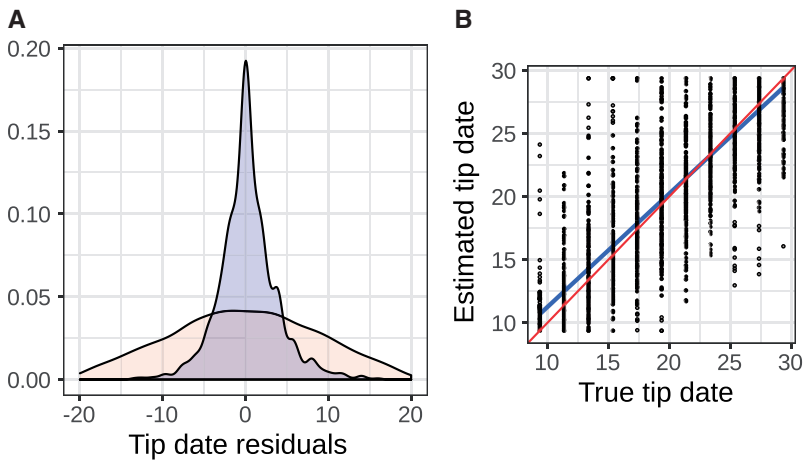

Figure 2. Inference of unknown time of sampling for 400 relaxed clock simulations described by To et al. (2015). A, Distribution of residuals of estimated and true tip dates. Red shows the residuals of the starting conditions and blue shows residuals after running the *treedater* algorithm. B, Estimated versus true tip dates. Blue shows a linear regression line and red shows the main diagonal.

The QPD algorithm (To et al. 2015) gave misleading results when temporal constraints were not enforced; the estimated TMRCA was later than the date of the first sequence (results not shown). When constraints were enforced, QPD estimated a TMRCA of 17 September 2012 rather than late 2013. We speculated that QPD may be giving different results because it is sensitive to outlier substitution rates in a small proportion of early samples, so we applied QPD to twenty phylogenies obtained by randomly downsampling to 250 tips and applying QPD to each subtree. QPD returned a mean TMRCA of 8 November 2013 across the twenty subtrees, similar to estimates with *node.dating*.

In order to compare the time-calibrated trees further, we applied skyspline (Volz, Romero-Severson, and Leitner 2017), a semi-parametric coalescent model that fixes the recovery rate and allows the number of new cases to vary over time, to the lineages-through-time for time-calibrated trees obtained using different methods. Figure 3 shows estimates of the cumulative number of infected cases over time, and Table 2 provides numerical summaries based on the total number of infections, and the timing of the peak of new cases per week, for which there are independent epidemiological estimates. Note that the epidemiological record is subject to unknown levels of under-reporting, and the true number of infections through time is not known. Also note that estimated number of cases will be sensitive to model structure, and the skyspline model assumes a simple susceptible-infected-recovered model with time-varying transmission rates. We find that skyspline applied to BRMC trees gives lower estimates for the number of cases, but provides an estimate of the peak that is consistent with epidemiological data. Skyspline applied to QPD gives estimates of the total number of cases that are very high, and peak too early. Skyspline applied to *treedater* trees gives estimates of the timing of the peak very similar to that obtained by BRMC, but with an estimate of the total number of cases that is closer to the number of reported cases. We were curious as to the drivers of the differences in magnitude of the number of infected cases obtained by BRMC and *treedater;* a notable difference in the BEAST MCC tree and the ML tree was the relatively high number of zero-length branches in the ML tree compared to the BEAST MCC tree (see Supplementary Figure S4). This difference arises due to the use of a prior on branch lengths in the BEAST phylogenetic reconstruction which smoothes these branches away from zero. To investigate the sensitivity of *treedater* to this phenomenon, we added a small number, equivalent to up to a single mutation, to either the tip lengths or the edge lengths of the ML tree, and reran *treedater*. Adding mutations to the tree resulted in much lower estimates of the number of cases, although the TMRCA and the timing of the peak number of cases changed relatively little. We also calculated the basic reproductive number, R_0 (operationally defined as the reproductive number at the TMRCA) using skyspline; all estimates were lower than those calculated from case onset data, although again, *treedater* gave point estimates that were similar to those obtained using the BEAST MCC tree.

4. Discussion

The *treedater* algorithm provides a new tool in a growing repertoire of software for molecular clock phylogenetic analysis, and fills a niche where existing tools may not provide acceptable performance. *treedater* is a fast method, like LSD and *node.dating*, and scales well to trees with thousands of lineages. While not as fast as LSD, *treedater* provides a flexible relaxed clock model of the substitution process that may be more realistic for many real data sets. *treedater* is integrated into the R statistical computing language and can be easily included in bioinformatic pipelines. There is substantial flexibility in the way *treedater* can be used; analyses may be run with or without rooted trees, with or without temporal constraints on nodes, and with strict or relaxed molecular clock models, in order to test sensitivity of results such as the effective population size to assumptions. We have added several capabilities to *treedater* that add to its utility for analysing biological datasets; 1, A PB approach, similar to the one implemented by To et al. (2015), provides confidence intervals for estimated substitution rates on each branch, the mean substitution rate, node dates, and lineages through time; 2, A statistical test based on the PB can be used to choose strict or relaxed molecular clock models (Kumar and Blair Hedges 2016; Duchêne et al. 2016); 3, The ability to accommodate missing tip dates, with arbitrary constraints for the times of sampling (compare to features in BEAST software, Drummond et al. 2006); and 4, The ability to identify outlier lineages, which may represent sequencing error or a different substitution process (compare to features in *Tempest* software, Rambaut et al. 2016).

Table 2. Point estimates of the TMRCA, the total number of cases, and the magnitude and timing of the peak number of new cases (per week) inferred from 1,610 Ebola virus sequences from the West African Ebola epidemic, collated by Dudas et al. (2017).

Method	Clock	TMRCA	Total cases	Peak of new cases (per week)	
Observed		December 2013	28,476	998 (28 November 2014)	1.71–2.02
BRMC	Relaxed	5 December 2013	15,697	423 (15 December 2014)	1.56
treedater	Relaxed	8 December 2013	48235	1284 (16 December 2014)	1.55
treedater + tips	Relaxed	3 December 2013	33,438	921 (14 December 2014-12-14)	1.59
treedater + edges	Relaxed	2 February 2014	24,871	627 (5 February 2015)	1.48
node dating	Strict	31 October 2013	29,514	730 (29 November 2014)	1.43
QPD	–	8 November 2013-11-08	201,660	5131 (18 December 2013)	1.37

Observed data refers to the number of reported cases in Guinea, Sierra Leone and Liberia from 1 March 2014 to 22 October 2015. The estimate of from the observed data is based on the time series of the number of cases, as estimates in WHO Ebola Response Team (2014). The point estimate of the TMRCA for BRMC is presented for the maximum clade credibility tree. The total number of infections, the magnitude and timing of the peak of new cases, and the basic reproductive number, are calculated by applying skyspline to each time-calibrated tree, with two spline points. All methods assumed temporal constraints on the tree. Note that QPD estimates are based on random subsampling of the ML tree (see text).

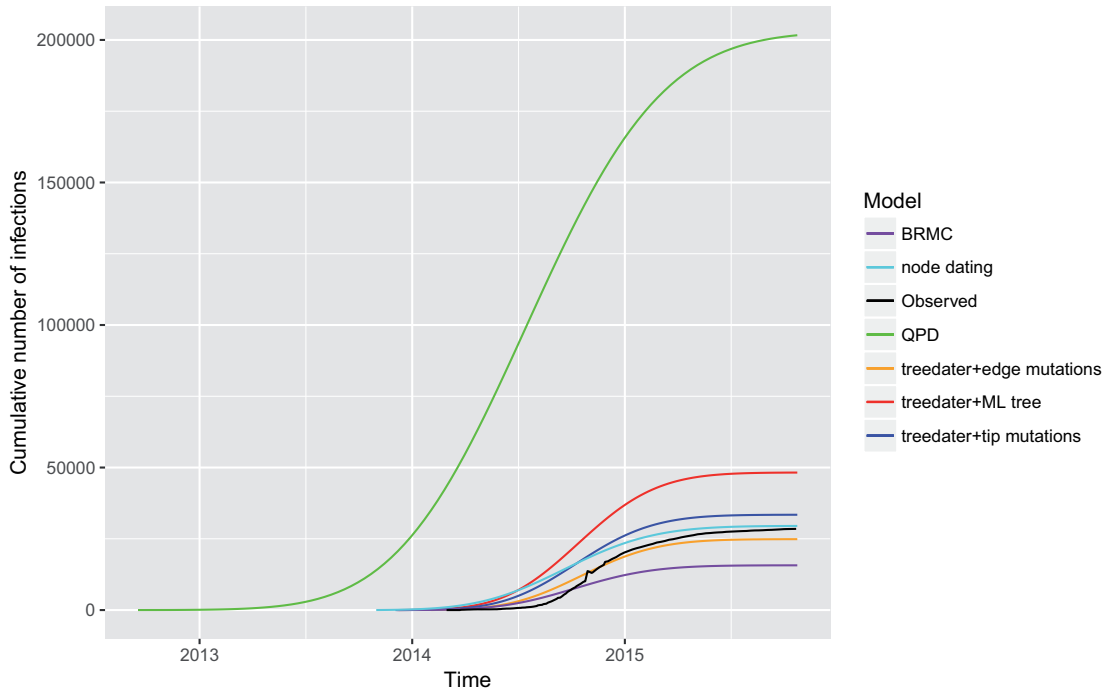

Figure 3. The estimated cumulative number of new cases of Ebola obtained by applying skyspline to rooted, time-calibrated trees obtained using different methods.

The iterative likelihood optimization procedure employed by *treedater* resembles commonly-used ML (expectation-maximization) and variational Bayes methods that are widely employed for difficult latent variable statistical models. This approach can be compared with the recently developed *node.dating* method. In the *node.dating* approach, most computational effort is expended on optimising the times of tree nodes given a mean substitution rate, which is treated as a nuisance parameter and typically estimated by fast RTT regression. In contrast, *treedater* treats the unobserved node dates as nuisance parameters, which are quickly estimated using a variation of the least squares algorithm presented by To et al. (2015) while conditioning on branch-specific substitution rates. Most computational effort in *treedater* is expended on optimising branch-specific substitution rates conditional on node dates. While the *treedater* algorithm relies on heuristic optimization, it is found to work surprisingly well in comparison to other methods focused on explicit optimization of a pseudo-likelihood (LSD) or sampling from a Bayesian posterior distribution (BEAST).

Application of *treedater* across a diverse range of simulations shows performance that is close to or superior to existing approaches across a wide range of scenarios with relatively low computational burden. When applied to a large dataset of Ebola virus sequences from the West African Ebola epidemic, *treedater* gives estimates of the time to the most recent common ancestor that are compatible with both epidemiological data and with more computationally intensive approaches such as those implemented in BEAST. In combination with skyspline, a high-throughput approach for inferring changes in population size over time from time-scaled phylogenies, *treedater* also gives estimates of the total number of cases and the timing and magnitude of the peak in new cases per week that are also compatible with epidemiological data.

There is substantial potential to further develop and extend *treedater*. Code optimization may bring speed and scalability close to LSD. Alternative models may allow substitution rates to be correlated between neighbouring branches (Gillespie 1984; Sanderson 2003) or to depend upon a population genetic model.

A statistical test could be developed to test for temporal signal in genetic data (Duchêne et al. 2015), and it may be possible to simultaneously estimate node dates and the parameters of a population genetic model such as the coalescent (Minin, Bloomquist, and Suchard 2008; Wakeley 2009) to estimate effective population size through time.

Funding

This study was supported by the National Institutes of General Medical Sciences, USA (U01GM110749 to EMV) and the Medical Research Council Centre for Outbreak Analysis and Modeling (MR/K010174 to EMV), a Methodology Research Programme grant from the Medical Research Council (MR/J013862/1 to SDWF) as well as by a supplemental grant to the Vanderbilt Center for AIDS Research, from the National Institute of Allergy and Infectious Diseases at the National Institutes of Health (AI110527).

Data availability

Code and data used for analysis of Ebola in Western Africa is available at https://github.com/sdwfrost/ebov-methods-comparison. Code and data used for simulation experiments is available at https://github.com/emvolz-phylodynamics/treedater-simulation-experiments.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Benjamini, Y., and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, 57: 289–300.

——, and Yekutieli, D. (2001) 'The control of the false discovery rate in multiple testing under dependency', *Annals of Statistics*, 29: 1165–88.

Biek, R. et al. (2007) 'A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus', *Proceedings of the National Academy of Sciences of the United States of America*, 104: 7993–8.

Blackley, D. J. et al. (2016) 'Reduced evolutionary rate in reemerged ebola virus transmission chains', *Science Advances*, 2/4: e1600378.

Drummond, A. et al. (2003) 'Inference of viral evolutionary rates from molecular sequences', *Advances in Parasitology*, 54: 331–58.

—— (2003) 'Measurably evolving populations', *Trends in Ecology & Evolution*, 18/9: 481–8.

—— (2006) 'Relaxed phylogenetics and dating with confidence', *PLoS Biol*, 4/5: e88.

Duchêne, S. et al. (2016) 'Estimating evolutionary rates using time-structured data: a general comparison of phylogenetic methods', *Bioinformatics*, 32(22): 3375–79.

—— (2015) 'The performance of the date-randomization test in phylogenetic analyses of time-structured virus data', *Molecular Biology and Evolution*, 32/7: 1895.

Dudas, G. et al. (2017) 'Virus genomes reveal factors that spread and sustained the ebola epidemic', *Nature*, 544/7650: 309–15.

Efron, B., and Tibshirani, R. (1994) *An Introduction to the Bootstrap*. London: CRC Press.

Gillespie, J. H. (1984) 'The molecular clock may be an episodic clock', *Proceedings of the National Academy of Sciences*, 81/24: 8009–13.

Greenwood, M., and Yule, G. U. (1920) 'An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents', *Journal of the Royal Statistical Society*, 83/2: 255–79.

Guindon, S. et al. (2010) 'New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0', *Systematic Biology*, 59/3: 307–21.

Holmes, E. C. et al. (2016) 'The evolution of ebola virus: Insights from the 2013–2016 epidemic', *Nature*, 538/7624: 193–200.

Jones, B. R., and Poon, A. F. (2017) 'node.dating: dating ancestors in phylogenetic trees in R', *Bioinformatics*, 33/6: 932–34.

Kumar, S., and Blair Hedges, S. (2016) 'Advances in time estimation methods for molecular data', *Molecular Biology and Evolution*, 33/4: 863.

Langley, C. H., and Fitch, W. M. (1974) 'An examination of the constancy of the rate of molecular evolution', *Journal of Molecular Evolution*, 3/3: 161–77.

McCloskey, R. (2015), ape::rtt. <https://github.com/cran/ape/blob/master/R/rtt.R> accessed 1 May 2017.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008) 'Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics', *Molecular Biology and Evolution*, 25/7: 1459–71.

Nguyen, L.-T. et al. (2015) 'Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular Biology and Evolution*, 32: 268–74.

Paradis, E., Claude, J., and Strimmer, K. (2004) 'APE: analyses of phylogenetics and evolution in R language', *Bioinformatics*, 20: 289–90.

Pybus, O. G. et al. (2003) 'The epidemiology and iatrogenic transmission of hepatitis c virus in egypt: a bayesian coalescent approach', *Molecular Biology and Evolution*, 20: 381–7.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A. et al. (2016) 'Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen)', *Virus Evolution*, 2/1: vew007.

Sanderson, M. J. (2003) 'r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock', *Bioinformatics*, 19/2: 301–2.

Stute, W., Manteiga, W. G., and Quindimil, M. P. (1993) 'Bootstrap based goodness-of-fit-tests', *Metrika*, 40/1: 243–56.

To, T.-H., et al. (2015) Fast dating using least-squares criteria and algorithms. *Systematic biology*, page syv068.

Turlach, Berwin A., and Weingessel A. quadprog: Functions to solve Quadratic Programming Problems., 2013. R package version 1.5-5.

Volz, E., Romero-Severson, M. E., and Leitner, T. (2017) 'Phylodynamic inference across epidemic scales', *Molecular Biology and Evolution*, 34/5: 1276–88.

Wakeley, J. (2009) *Coalescent Theory*. Greenwood Village, Colorado, USA: Roberts & Company.

WHO Ebola Response Team (2014) 'Ebola virus disease in west africathe first 9 months of the epidemic and forward projections', *N Engl J Med*, 371: 1481–95.

Wood, S. N. (2006) *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC. London: Chapman & Hall/CRC.